International Journal of Electrical and Computer Engineering (IJECE)
Vol. 4, No. 1, Feburary 2014, pp. 37~51
ISSN: 2088-8708

A Novel Semantic Statistical Model for Automatic Image Annotation Using the Relationship between the Regions Based on Multi-Criteria Decision Making

Hengame Deljooi[1], Ahmad R. Eskandari[2]
[1]Department of Electrical, Computer and IT engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[2]Digital Media Lab, AICTC Research Center, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Article Info

Article history:

Received Sep 30, 2013
Revised Dec 23, 2013
Accepted Jan 10, 2013

Keyword:

Automatic Image Annotation
Multi Criteria Decision Making
Regional Context
Statistical Models
Visual Topic

ABSTRACT

Automatic image annotation has emerged as an important research topic due to the existence of the semantic gap and in addition to its potential application on image retrieval and management. In this paper we present an approach which combines regional contexts and visual topics to automatic image annotation. Regional contexts model the relationship between the regions, whereas visual topics provide the global distribution of topics over an image. Conventional image annotation methods neglected the relationship between the regions in an image, while these regions are exactly explanation of the image semantics, therefore considering the relationship between them are helpful to annotate the images. The proposed model extracts regional contexts and visual topics from the image, and incorporates them by MCDM (Multi Criteria Decision Making) approach based on TOPSIS (Technique for Order Preference by Similarity to the Ideal Solution) method. Regional contexts and visual topics are learned by PLSA (Probability Latent Semantic Analysis) from the training data. The experiments on 5k Corel images show that integrating these two kinds of information is beneficial to image annotation.

Corresponding Author:

Ahmad R. Eskandari
Digital Media Lab, AICTC Research Center
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Email: h.deljooi@qiau.ac.ir, eskandari@dml.ir

1. INTRODUCTION

With the advent of digital imaging and data storage, searching and indexing large image databases efficiently and effectively has become a challenging problem [1-5]. In order to solve the problem, there are three distinct approaches coexist in the literature [1-3].

The first approach is the traditional Text Based Image Retrieval (TBIR). In this approach images are annotated manually with keywords or captions and then retrieved using a conventional text search engine. This technique uses text to capture semantic content of images and allows query by text. However, expensive labor makes this solution difficult to be extended to large image databases. Furthermore, human annotations are usually too subjective and ambiguous.The second type is Content Based Image Retrieval (CBIR), in which queries are usually based on visual examples. In CBIR various low-level visual features are extracted from each image in the database and image retrieval is formulated as searching for the best database match to the feature vector extracted from the query image. Although this process is accomplished quickly and automatically, the results are seldom semantically relevant to the query example due to the notorious semantic gap [6]. The semantic gap highlights the wide difference between semantic concept used by humans to interpret images and low level features.The third approach is Semantic Based Image Annotation (SBIR).

To bridge the semantic gap, Automatic Image Annotation (AIA) techniques have attracted a lot of interest in SBIR systems.

Automatic image annotation is a process to automatically generate textual words to describe the content of untagged images from annotated images according to image similarities. The goal of automatic image annotation is to find the relation between low level visual features and high level semantic concepts. A main problem in automatic image annotation is to create a model to assign keywords to an image in order to describe it. In this regard it is necessary to train a set of images have already been annotated by humans manually. These annotations are semantic concepts made up of simple keywords that described the content of the image. Image analysis techniques are used to extract features from the images such as color, texture and shape in order to model the distribution partitions of the image. The next step is to extract the same feature information from an unseen image in order to compare it with all previously created models (one for each keyword). The result of this comparison yields a probability value for each keyword included in the image..

Literature review that has been done author used in the chapter "Introduction" to explain the difference of the manuscript with other papers, that it is innovative, it are used in the chapter "Research Method" to describe the step of research and used in the chapter "Results and Discussion" to support the analysis of the results [2]. If the manuscript was written really have high originality, which proposed a new method or algorithm, the additional chapter after the "Introduction" chapter and before the "Research Method" chapter can be added to explain briefly the theory and/or the proposed method/algorithm [4].

1.1. Relatedworks

In recent years, many algorithms has introduced for AIA. We can roughly classify them into 4 categories: **The Vector Space Models**: The vector space model [7,8] framework is a popular technique in information retrieval especially in text retrieval. Generally, documents are represented as vectors, each of which contains the occurrences of keywords within the document in question. The length of the vectors is equal to the vocabulary size. These approaches treat images as documents, and build visual terms which are analogous to keywords, from the image feature descriptors. Some of the methods in this category include: SvdCos method, Saliency-based semantic propagation, and cross-language latent semantic indexing based approach. **Classification Methods**: Classification approaches [9-11] for image annotation view the process of attaching keywords to images as that of classifying images to a number of pre-defined groups, each of which is characterized by a concept or keyword. These approaches pose image annotation as a supervised classification problem. Specifically, each keyword is viewed as a unique class. Binary classifiers for each class or a multiclass classifier is trained independently to predict the annotations of new images. Given an un-annotated image, the classification algorithms find its membership and annotate it with the corresponding keyword. Multiple annotations can be generated by assuming an image belongs to multiple classes. Some of the works in this category include: image linguistic indexing, image annotation using SVM, multiple instance learning approaches, non-negative Matrix factorization based approaches. **Graph Based Methods**: Recently, the graph-based methods [12-17] have achieved much success in the image and video analysis domain including image annotation. How to build a similarity graph is very important in graph learning. A good graph should reflect a deep understanding of the data structure and help to mine potential knowledge as much as possible. Some works in this category include: image annotation via graph learning, image annotation refinement via graph based learning. **Statistical Models**: The basic idea of statistical techniques [18-26] is to estimate the probabilities of documents related to the query submitted by the user. Documents are then ranked according to their probabilities. The main goal of statistical techniques is to learn a model from the joint distribution of visual and textual features on the training data and predicting the missing textual features for a new image.

Among the image annotation models classification approaches are the oldest in this domain. But statistical models and graph based approaches have more chance. The proposed approach in this paper follows statistical based models. So we will review the statistical models that are proposed for automatic image annotation.The co-occurrence model proposed by Mori et al. [20] is perhaps one of the first attempts at automatic image annotation. They first divide images into rectangular tiles of the same size, and calculate a feature descriptor of color and texture for each tile. All the descriptors are clustered into a number of groups, each of which is represented by the centroid. Each tile inherits the whole set of labels from the original image. Then, they estimate the probability of a keyword W related to a cluster C by the co-occurrence of the keyword and the image tiles within the cluster.Duygulu et al. [21] proposed a machine translation model for automatic image annotation. They argued that region based image annotation is more interesting because global annotation does not give information on which part of the image is related to which keyword. In their point of view, the process of attaching keywords to image regions is analogous to the translation of one form of representation (image regions; French) to another form (words; English). Indeed, Their Machine translation model applies one of the classical statistical machine translation models to translate from the set

of keywords of an image to the set of blobs that generated by clustering the image features. Jeon et al. [22] proposed Cross Media Relevance Model (CMRM) for automatic image annotation. CMRM benefits the joint distribution of keywords and blobs. Each image describe by visual features (blobs) as well as textual keywords. In CMRM the probability of observing a set of blobs with a set of keywords estimated. Lavrenko et al. [23] proposed a Continues-space Relevance Model (CRM) that improves CMRM by using continuous probability density functions to estimate the probability of observing a region given an image. Feng et al. [24] modified the above model and proposed Multiple Bernoulli Relevance Model (MBRM) such that the probability of observing labels given an image was modeled as a multiple-Bernoulli distribution. In MBRM images simply divided into rectangular tiles instead of applying automatic segmentation algorithms.

There exist two problems in above algorithms. First, most existing algorithms have taken one of two approaches, using regional or global features exclusively. Second, conventional approaches consider each word separately without the textual context relations. As a result the textual context relations among annotation words have been ignored. By textual contextwe refer to co-occurrence relationship among words. Wang et al. [25] proposed a model for image annotation that combines global, regional, and contextual features by an extended CMRM. They incorporate the three kinds of information to describe image semantics to annotate images by estimating their joint probability. Some of the other statistical based models for image annotation include: D-CMRM, SDF and Probabilistic semantic model.

1.2. Ourapproach

There exists an important problem in all of the statistical methods, that all the explained methods utilize direct distribution of regions for AIA. However considering the relationship between regions, that each of them is an explanation of a word, helps us to improve the final annotation words. So we obtain the relationship between regions and instead of using the direct distribution of regions, we utilize the distribution of topics between regions for AIA.

To address the above problem we propose a novel approach for AIA. Typical CMRM only takes regional features to describe an image, the extended CMRM incorporates both global and regional features, as well as textual context to annotate images. In the approach that we propose, the *visual topics* have described as a global distribution vector of topics in the image, moreover we consider the obtained relationship between regions and model the *regional contexts* as a distribution of topics between regions. Both the visual topics and regional contexts are learned by a PLSA approach [27] from the training data. Then we integrate these two kinds of information by MCDM approach based on TOPSIS method [28,29]. Generally our method includes three steps as follows:

1. Model the regional contexts or the relationship between the regions as a distribution of topics between regions for generating the annotation related to image regions content.
2. Utilize the visual topics as a global distribution vector of topics to consider the overall image content.
3. Combining the extracted two kinds of information from the image using the TOPSIS method in MCDM algorithm.

The framework is shown in Figure 1. The images are represented by their regions and collection of patches. Our proposed model is learned from the training data based on the two kinds of information, regional contexts $B(I)$ and visual topics $H(I)$.

1.3. Paperoutline

The reminder of this paper organized as follows: Extended CMRM is described in section 2. Our proposed model is introduced in section 3. In section 4 experimental results are presented and finally concluding remarks are explained in section 5.

2. EXTENTED CMRM

The conventional CMRM [22] considers only one representation of images, i.e. a bag of blobs. To deal with images which are not suitable to be represented as a bag of blobs, we need to consider other representation as well. So, ECMRM [25] has proposed to use visual topics as another representation. This new image representation is combined with the visual blobs representation. Moreover the typical CMRM annotates keywords individually without considering the joint distribution of different keywords. To solve this problem, textual context is proposed. To annotate an image with multiple keywords, first annotate the image with textual contexts and then compose the keywords from the distribution of keywords under each textual context. The learning of textual context is based on the PLSA approach [27]. PLSA is proposed to automatically learn topics from text documents. Similar to learn topics from text documents, we can also learn visual topics from images.

Similar to learning topics from text documents, we can also learn visual topics from a collection of images. The main point is representing an image as a bag of words [30] similar to the vector representation of text documents. In details, we partition an image by a regular grid and take it as an unordered set of image patches. Then we extract a 128-D SIFT descriptor [31] and vector-quantize each image patch by clustering [32] a subset of patches from the training images, which has proved effective for object recognition [33]. We call the set of cluster centers as visual vocabulary. We can then transform an image into a bag of visual words by assigning a visual word label to each image patch. Given this bag of visual words representation, it is then straightforward to apply PLSA to learn a set of visual topics, each of which is characterized by a multinomial distribution of visual words.The ECMRM method annotates a test image *I* by estimating the joint probability of textual contexts *C*, its visual blobs (regions) $R=(b_1,b_2,...,b_m)$ obtained by image segmentation [34] and the visual topics distribution $H(I)$, which *J* represents training image and τ is the training set:

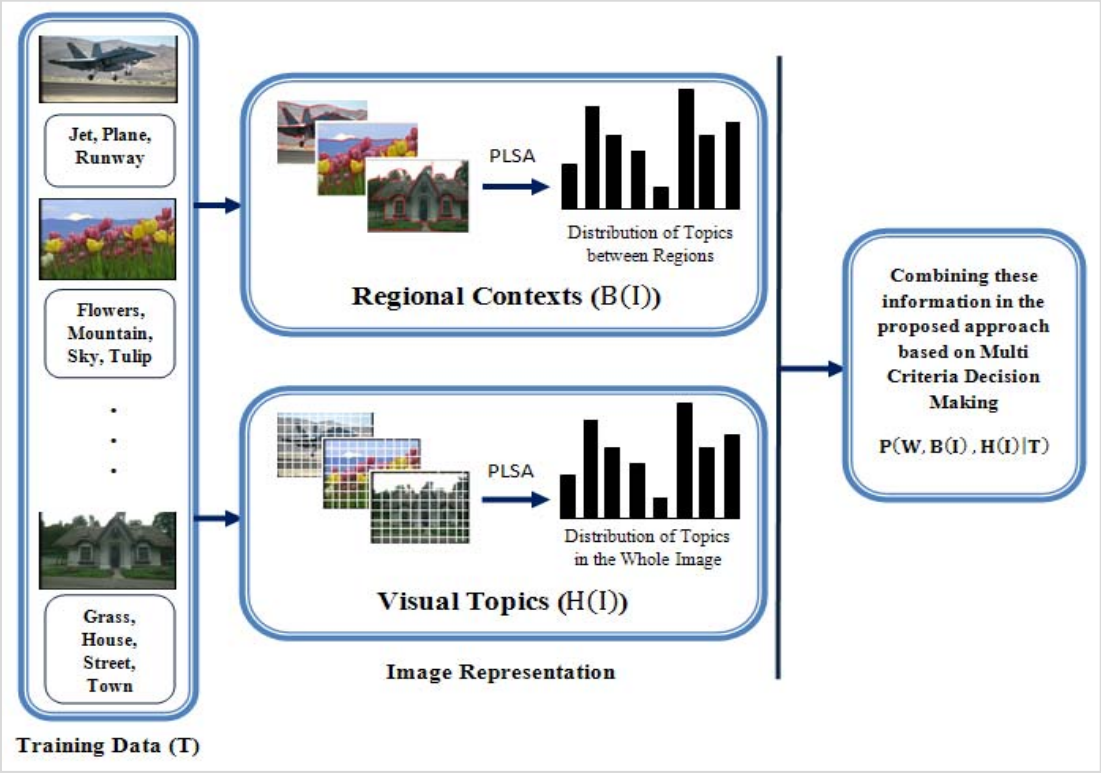

Figure 1. The proposed approach to automatic image annotation using regional contexts and visual topics.

$$P(C,R,H(I))=\sum_{J\in\tau}P(J)P(C,b_1,...,b_m,H(I)|J) \qquad (1)$$

Comparing ECMRM with CMRM, there are two points of difference to elaborate. First, the original CMRM in annotates an image using only the regional features $R=(b_1,b_2,...,b_m)$. However, ECMRM uses both the regional features *R* and the global features $H(I)$ which represent the global distribution of visual topics in image *I*. This suggests the ECMRM model combines the global features and regional features. Second, CMRM predicts the probability of a single word *W* directly, while ECMRM predicts the probability of a textual context *C*. This indicates ECMRM does not assume the mutual independence between words given the image. Thus, the extended CMRM incorporates the textual context from the training data.

Textual contexts *C*, image blobs and visual topics distribution are independent. So that $P(C,b_1,...,b_m,H(I)|J)$ can be simplified as:

$$P(C,b_1,...,b_m,H(I)|J)=P(C|J)P(H(I)|J)\prod_{i=1}^{m}P(b_i|J) \qquad (2)$$

$P(b|J)$ is estimated as the same as that in CMRM. $P(C|I)$ is available after learning textual contexts on the manual annotations. $P(H(I)|J)$ is defined as the Kullback-Leibler divergence [35] between the visual topic distribution of *I* and *J*:

$$P(H(I)|J) = D_{kl}(H(I)|H(J)) = \sum_{i=1}^{Q} P(q_i|I) \log \frac{P(q_i|I)}{P(q_i|J)} \quad (3)$$

In which D_{kl} is the Kullback-Leibler divergence between two distributions. From the Bayesian theory, we know that:

$$P(C|I) = \frac{P(C,I)}{P(I)} = \frac{P(C,b_1,...,b_m,H(I))}{P(I)} \quad (4)$$

Therefore, normalization on $P(C,b_1,...,b_m,H(I))$ will give the conditional distribution of textual contexts $P(C|I)$. The conditional keyword distribution $P(W_j|I)$ of *I* is obtained by fusing the keyword distribution of the entire textual contexts:

$$P(W_j|I) = \sum_{i}^{S} P(W_j|C_i) P(C_i|I) \quad (5)$$

3. THE PROPOSED APPROACH

The notable point is that the relationship between regions is not considered in the ECMRM method. We know that each of these regions is explanation of a word and the semantic relationship between these regions is more accurate than the semantic relationship between words (textual contexts), which are assigned to the image by human. Because these regions have represented the exact content of an image, finally the generated keywords are related to the image content.

3.1. Learning Regional Contexts from Images

Similar to learning topics from text document in the ECMRM method, we can also learn regional contexts from a collection of images. Firstly, according to ECMRM method the regions are obtained from image segmentation based on objects [34]. We should represent an image as a collection of words, so consider an image as a document and assume the objects in an image as words in this document (we consider each region as a word). Then apply the PLSA approach developed by Hoffman to reach the relationship between regions and instead of the direct distribution of regions in an ECMRM, we utilize the distribution of topics between them to AIA. We call these topics, which have described the relationship between regions, regional context.

PLSA approach [27] behaves as follows:

Suppose we are given a set of text documents $D = \{d_1, d_2, ..., d_n\}$ each of which is represented by a term frequency vector:

$$d_i = [n(d_i, w_1), n(d_i, w_2), ..., n(d_i, w_m)] \quad (6)$$

In which $n(d_i, w_j)$ is the number of occurrence of word w_j in document d_i, and *m* is the vocabulary size. PLSA assumes that each word in a document is generated by a specific hidden topic Z_k, where $Z_k \in I$ and *I* is the vocabulary of hidden topics. Since Z_k is a hidden variable, the conditional probability of a word w_j given document d_i is a marginalization over the topics:

$$P(w_j|d_i) = \sum_{k}^{K} P(w_j|Z_k, d_i) P(Z_k|d_i) \quad (7)$$

In which K is the number of hidden topics, $P(w_j|Z_k, d_i)$ is the conditional probability of a word w_j given topic Z_k and the document d_i, $P(Z_k|d_i)$ is the conditional probability of topic Z_k given d_i. Furthermore, PLSA assumes that the conditional probability of generating a word by a specific topic is independent from the document:

$$P(w_j|Z_k, d_i) = P(w_j|Z_k) \quad (8)$$

Therefore (7) can be simplified as:

$$P(w_j|d_i) = \sum_{k}^{K} P(w_j|Z_k) P(Z_k|d_i) \quad (9)$$

The model parameters $P(w_j|Z_k)$ and $P(Z_k|d_i)$ can be learned by an EM algorithm [36].

3.2. Learning Visual Topics from Images

Similar to learning topics from text documents, we can also learn visual topics from images. The notable point is representing an image as a bag of words [30], like to the vector representation of text documents. We partition an image by a regular grid and take it as an unordered set of image patches. Then we extract a Gabor descriptor [37-39] and vector quantizes each image patch by clustering [32] a subset of patches from the training images, which has proved effective for object recognition [33]. Therefore, we develop the orientation histogram [40,41] from multi-scale Gabor features because Gabor features are better representation than simple gradient features. We call the cluster centers as visual words. Then we can transform an image into a bag of visual words by assigning a visual word label to each image patch. We have the bag of visual word representation, apply the PLSA to learn a set of visual topics.

There is a difference between image regions obtained by image segmentation and image patches. Image patches that grouped in a topic do not have spatial agglomeration and visual consistency although image segmentation groups pixels by its visual property and spatial location. Figure 2 illustrates the image regions, image patches and the distributions of regional context and visual topics of an image. Visual topics and regional contexts (topics between regions obtained by image segmentation) focus on different aspect of an image, so they are complementary to each other and a combination of them is expected to achieve better performance.

3.3. Combining Regional Contexts and Visual Topics

Our method annotates a test image I by combining the regional contexts (the distribution of topics between regions) $B(I)$ and the visual topics distribution $H(I)$ and considering a given word (W):

$$P(W, B(I), H(I)) = \sum_{J \in \tau} P(J) P(W, B(I), H(I)|J) \quad (10)$$

Comparing (10) with (1) there is one point of difference to elaborate: ECMRM in (1) annotates an image uses both the regional features R and the global features $H(I)$, means that it uses the direct distribution of regions. But our method in (10) considers to the relationship between the regions and it uses the distribution of topics between regions as a regional contexts $B(I)$. Because the semantic relationship between regions is more accurate than the relationship between words, we have ignored the textual context in (1) and instead of it we utilize regional contexts. So we utilize both the regional contexts and visual topics together. These two kinds of information extract different words, so combining them are useful to create related words and consistent to image content. However the conventional methods suppose the conditionally independence between the regional information and visual words, but we cannot consider this assumption. To describe this point the dependency graph of the image, visual words and regional information is shown in the Figure 3. It

is obvious that to obtain the regional information we cannot utilize the image directly and we should use the image features domain. Subsequently regional information is dependent to the visual words, because the visual words are in this domain too. This dependency stems from that one of the important extracted features from the image needed for regional representation is texture which cannot be independent from the visual words.

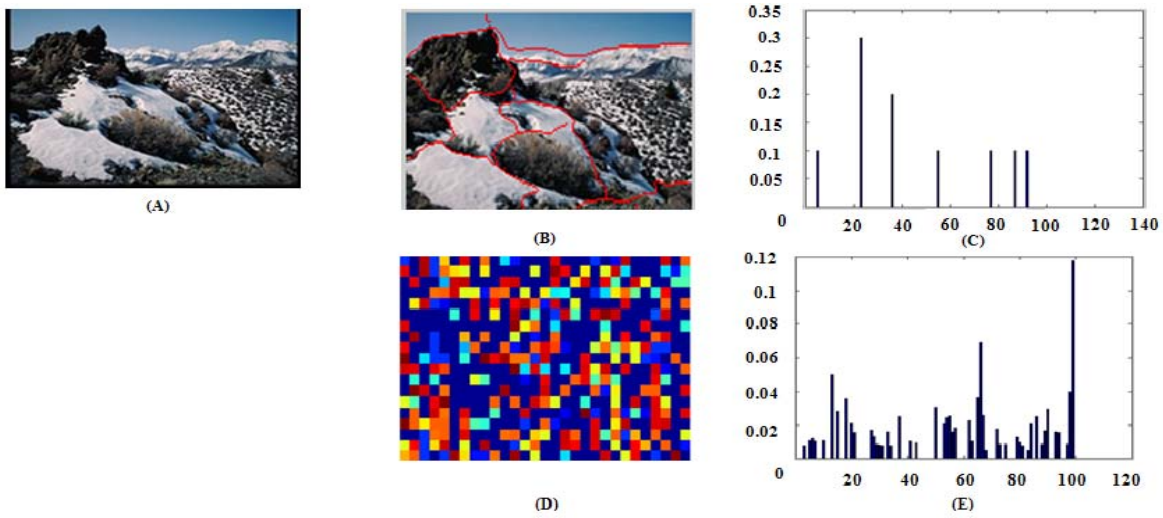

Figure 2. Illustration of two types of information: (A) original image, (B) image regions have obtained by segmentation, (C) distribution of regional contexts, (D) Image patches that the patches with the same topic are indicated by the same color, (E) distribution of visual topics.

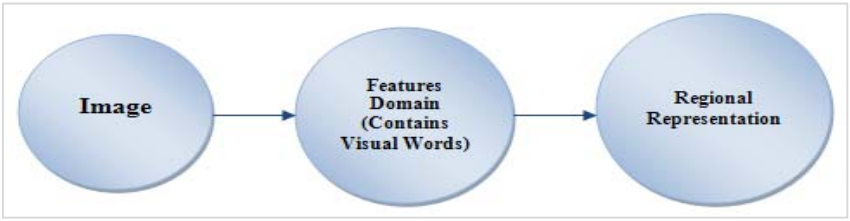

Figure 3. Graph dependency of the image, visual words and regional information.

To solve this problem we explain the distribution of test image keywords generally as below equation:

$$p(w|I) = \sum_{J \in \tau} p(w|J)\alpha_j \qquad (10)$$

This equation is similar to (10), but we utilize α_j instead of considering the joint probability of $B(I)$ and $H(I)$. According to (11) the annotation of a test image is equal to the sum of multiply the distribution of training image keywords in their coefficient (α_j). This coefficient calculate the similarity of a test image *I* with the training images, considering to (10) in our proposed model α_j is the conditional joint distribution of the regional contexts and visual topics. Indeed α_j define the contribution of training image keywords to create the distribution of test image keywords. We can understand easily that if one training image is more similar to the test image *I*, the distribution of their keywords is similar too, and the contribution of that training image or α_j should be more. As regards $B(I)$ and $H(I)$ are dependent and to simplify (10) and calculate α_j or the similarity between the test image and training images we use the

MCDM (Multi Criteria Decision Making) approach [28]. MCDM is the selection of the best action from a set of alternatives, each of which is evaluated against multiple criteria. This method is commonly used in fields different from structural engineering, for example for the resources allocation planning, for ranking the sequential patterns and to locate a special facility. Also MCDM was applied to financial investment in advanced manufacturing systems.

Among the MCDM methods, we utilize the TOPSIS (Technique for Order Preference by Similarity to the Ideal Solution) method [28, 29]. Some of the advantages of this algorithm is the ability of determining any kinds of criteria, the explicitly of its results and decrement the complexity of the parameters. Moreover this method has been used to calculate the similarity. In this method two artificial alternatives are hypothesized, Ideal alternative: the one which has the best level for all attributes considered.Negative ideal alternative: the one which has the worst attribute values. TOPSIS evaluates the alternatives based on the closeness to the ideal solutions and farness from negative ideal solutions.

Inputs to TOPSIS methods are as follows:

- TOPSIS assumes that we have m alternatives (options) and n attributes (criteria) and we have the score of each option with respect to each criterion. In our model the alternatives are training images and the criteria are regional contexts and visual topics. Indeed to determine the criteria we should introduce some measure to calculate the similarity. We obtain these measures by investigating the various experimental results. In the different experiment, we fetch up that $\left[p(B(I))^2\right]$ and $\left[p(B(I))\right]$ increase the precision and $\left[p(B(I))*p(H(I))\right]$ has an influence to boost the recall, so we specify these measures:

$$\left[p(B(I)|J)^2, p(B(I)|J), p(B(I)|J)p(H(I)|J)\right] \qquad (12)$$

- Let x_{ij} score of option *i* with respect to criterion *j*, we have the m×n matrix.
- Let *J* be the set of benefit attributes or criteria (more is better).
- Let *J'* be the set of negative attributes or criteria (less is better).

Step 1: Construct normalized decision matrix.

- This step transforms various attribute dimensions into non-dimensional attributes, which allows comparisons across criteria.
- Normalize scores or data as follows:

$$r_{ij} = \frac{x_{ij}}{\sqrt{x_{ij}^2}} \quad \text{for i=1,...,m:j=1,...,n} \qquad (13)$$

Step 2: Construct the weighted normalized decision matrix.

- Assume we have a set of weights for each criterion w_j for $j=1,...,n$. These weights express the relative importance of it in respect to the others. Considering the 3 suggested criteria in our model, we need to define 3 weights for them:

$$\left[a*\frac{6}{7} \quad \frac{a}{7} \quad 1\text{-}a\right] \qquad (14)$$

- Multiply each column of the normalized decision matrix by its associated weight.
- An element of the new matrix is:

$$v_{ij} = w_{ij} \cdot r_{ij} \qquad (15)$$

Step 3: Determine the ideal and negative ideal solutions for each criteria.

- Ideal solution:

$$A^* = \left\{v_1^*, ..., v_n^*\right\} \text{ where } v_j^* = \left\{\max\left(v_{ij}\right) \text{ if } j \in J : \min\left(v_{ij}\right) \text{ if } j \in J'\right\} \qquad (16)$$

- Negative ideal solution:

$$A' = \left\{ v_1', ..., v_n' \right\} \text{ where } v' = \left\{ \min\left(v_{ij}\right) \text{ if } j \in J : \max\left(v_{ij}\right) \text{ if } j \in J' \right\} \quad (17)$$

Step 4: Calculate the separation measures for each alternative (training images). The separation (distance) between alternatives can be measured by the n dimensional Euclidean distance.

- The separation from the ideal alternative is:

$$S_i^* = \left[\sum_j \left(v_j^* - v_{ij} \right)^2 \right]^{\frac{1}{2}} \quad i=1,...,m \quad (18)$$

- Similarly, the separation from the negative ideal alternative is:

$$S_i' = \left[\sum_j \left(v_j' - v_{ij} \right)^2 \right]^{\frac{1}{2}} \quad i=1,...,m \quad (19)$$

Step 5: Calculate C_i^* or similarities to the ideal solution (α_j).

$$C_i^* = \frac{S_i'}{\left(S_i^* + S_i' \right)} \quad 0 < C_i^* < 1 \quad (20)$$

Note that original TOPSIS chooses an alternative with C_i^* closest to 1, but in our model we need all the value of C_i^* for image annotation and generate the keywords of test image, because these values are α_j or the coefficient of the training images.

Considering to the suggested criteria, we need to calculate the $P\left(B(I)\middle|J\right)$ and $P\left(H(I)\middle|J\right)$. These two items are defined as the Bhattacharyya metric distance[42] between the distribution image *I* and image *J*:

$$P\left(B(I)\middle|J\right) = D\left(B(I)\middle|B(J)\right) = \sqrt{1 - \sum_i \sqrt{p_i q_i}} \quad (21)$$

$$P\left(H(I)\middle|J\right) = D\left(H(I)\middle|H(J)\right) = \sqrt{1 - \sum_i \sqrt{p_i q_i}} \quad (22)$$

In which *D* is the Bhattacharyya metric distance between two distributions. In the Bhattacharyya metric distance, small distances between two distributions show small weight values and vice versa, consequently it's better to apply the following equation in *D* to achieve a good result:

$$P = \frac{1}{\sqrt{2\pi\sigma}} \left(\exp\left(-\frac{D^2}{2\sigma} \right) \right) \quad (23)$$

In the above mentioned equation, we observe that small distances correspond to big weight values, while big distances (not correspondence) generated small weight values. That is specified by a Gaussian with variance σ and we can improve the result with tuning σ. Further in this equation *D* is the Bhattacharyya metric.

4. RESULTS

In this section we will discuss details of the dataset and also show experimental results. The test bed and evaluation metrics are introduced in Section 4.1 and Section 4.2, respectively. The experimental results with compared to other annotation models are presented in Section 4.3

4.1. Testbed

We have used a set of Corel images consists of 5000 images to evaluate the proposed model. The image database consists of 50 main topics and each class includes 100 images. COREL-5K was first collected by Duygulu et al. [21] and it has been extensively used by other researchers, so it has been known as a de facto standard dataset in image annotation researches. Each image is also associated with 1-5 keywords. Therefore, there are 374 keywords in the dataset.

We partition the whole data set into a training set and a test set, 4500 training images and 500 test images. For the region features, we use the JSEG algorithm [34] to segment each image into 1--11 regions. Image regions with area less than 1/25 of the whole image are discarded. In average there are five image regions per image. Each image region is represented by a 36-D feature vector. For the dense grid, we sample 13×13 pixels image patches without overlapping. The average number of image patches per image is around 550. The image regions are clustered into 500 image blobs, similarly the Gabor descriptor of image patches are clustered into 500 centered. We experiment on different number of visual topics (V), textual context (T) and regional contexts (R) and achieve the best number of them that are shown in results.

4.2. Evaluation Metrics

For each method, we take the top five words as the final annotation. We use precision, recall and F_1 metrics, to evaluate the quality of our approach. Precision and recall are calculated on the basis of each keyword in the vocabulary. Then the values are averaged over the keywords in the vocabulary.

Precision is defined as the number of images correctly predicted with a given keyword called *r*, divided by the total number of images predicted with the keyword called *n*. Recall is defined as the number of images correctly predicted with a given keyword, divided by the total number of images having this keyword in its ground-truth called *N*.

$$\Pr ecision(W) = \frac{r}{n} \ , \ \mathrm{Re}call(W) = \frac{r}{N} \ , \ F_1 = \frac{2 \times \Pr ecision \times \mathrm{Re}call}{(\Pr ecision + \mathrm{Re}call)} \qquad (24)$$

We also use keyword number with recall>0 to show the diversity of correct words that can be predicted by the automatic image annotation model. The measure is important because a model can achieve high precision and recall values by only performing well on a small number of common words.

4.3. Comparativeanalysis

In this section, we present a comparative analysis among some research work focusing on different parts of the proposed model. We design two types of experiments, evaluate the effect of applying regional contexts in image annotation and then analyze combining this information with visual topics based on MCDM in annotation process. Finally we present an overall comparison with previous annotation models.

4.3.1. Comparison on effect of regional contexts

In the first experiment we evaluate the effect of regional context, discussed in section 3.1. The results of experiments are shown in table 1. Comparing the results in a table, the improvement of ECMRM [25] which incorporates the regional, global and contextual features to original CMRM [22] is obvious and improved value from 0.1115 to 0.3270. But in the proposed approach using the regional contexts improved value from 0.3270 to 0.4214. In addition precision, recall and keyword number with recall>0 measure have been significantly improved. In summery using the regional contexts (neglect the textual context), significantly improve the performance, because the regions or objects in an image are exactly the explanation of image content. Also each of these regions acts as one word, so the relationship between the regions is more accurate than textual contexts to create annotation keywords. Textual contexts have some noises, because the keywords of training images are created manually and by human, consequently they have inconsistence.

4.3.2. Comparison on combining regional contexts and visual topics

In this experiment we investigate the performance of our model consists of integrating regional contexts and visual topics based on MCDM, discussed in section 3.2 and 3.3. The results of experiments are

shown in table 1 too. The effect of regional contexts to improve the performance is shown in the previous experiment. It is obvious that incorporating this information with visual topics help us to achieve the best performance, because these two kinds of information focus on different aspects of an image, extract different words and they are complementary to each other. This improvement is shown in table 1 and increase the F_1 value from 0.4214 to 0.4938. Fig. 4 shows some sample words and their precision and recall values and Fig. 5 shows F_1 value of sample words in CMRM, ECMRM and the proposed approach. The difference between the F_1 value of our method and ECMRM of sample words in Fig. 5 shows the noises in the manual annotation and textual contexts of ECMRM method. We can observe that considering the relationship between regions is helpful to obtain the annotation keywords related to image content and increase the F_1 value.

4.3.3. Overallcomparison

In the both experiments mentioned above, performance evaluation of the proposed model has been illustrated from viewpoint of using regional contexts and viewpoint of integrating regional contexts and visual topics based on MCDM to image annotation. In this section the performance of our model (combining regional contexts and visual topics based on MCDM) is compared with other related works. The overall comparison (the average precision and recall) in some statistical models is illustrated in Fig. 6 and the number of keyword number with recall>0 in that statistical models is shown in Fig. 7, that the predominance of our model compared to other models is obvious. Some test images with the annotations generated by the CMRM, ECMRM and the proposed approach are demonstrated in Fig. 8. It is observed that the proposed approach can yield better annotations, which describe image semantics, than ECMRM.

Table 1. The average precision, recall and f_1 values of experiments

Models	CMRM [22]	ECMRM [25]	The proposed approach using regional contexts	The proposed approach using combination of regional contexts and visual topics
Precision	0.1288	0.3050	0.4408	0.5186
Recall	0.0983	0.3524	0.4038	0.4712
F_1	0.1115	0.3270	0.4214	0.4938
Keywords with recall>0	66	129	133	144

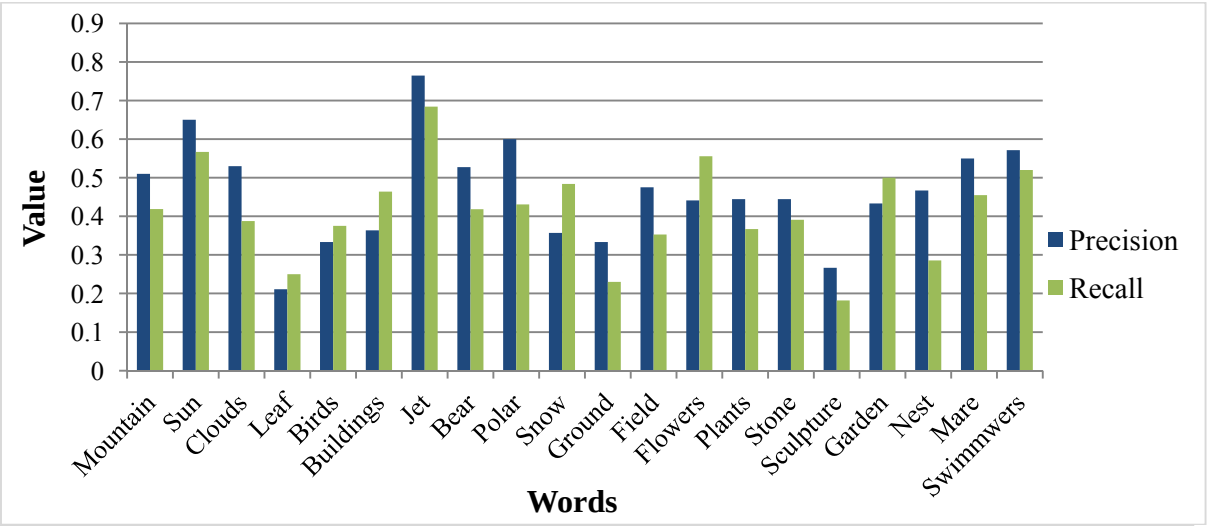

Figure 4. Some sample words and their precision and recall values in the proposed approach

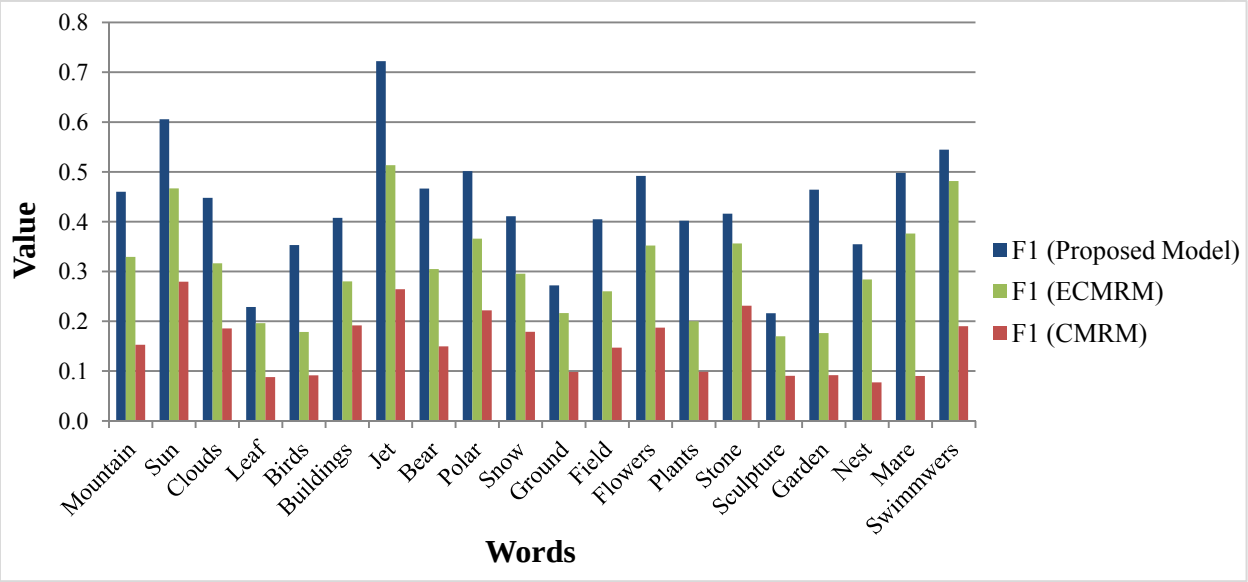

Figure 5. Some sample words and their F_1 value in CMRM, ECMRM and the proposed approach

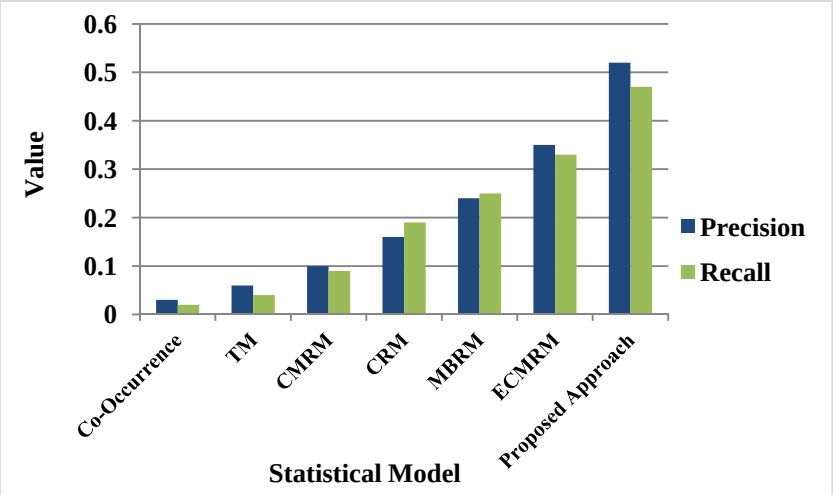

Figure 6. The average precision and recall in some statistical models

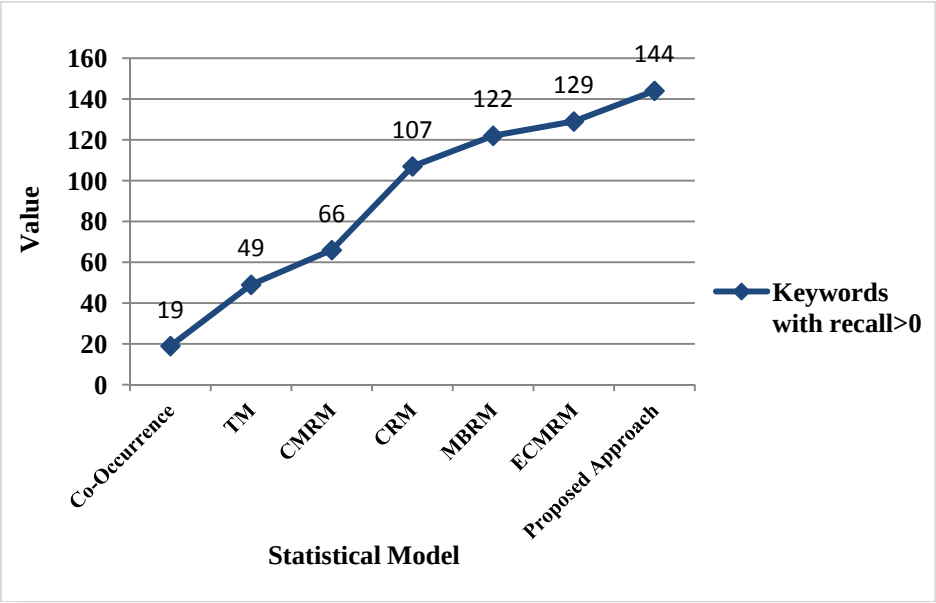

Figure 7. Kewords with recall>0 in some statistical models

Image	Ground Truth	CMRM	ECMRM	Proposed Model
	Mountain, Sky, Sun, Water	Building, Sunset, Town, Water, Tree	Building, Sunset, Water, People, Tree	Tree, Sky, Clouds, Sunset, Water
	Clouds, Grass, Mountain, Town	Plane, Sky, Jet, Water, Grass	Water, Plane, Sky, Grass, Tree	Sky, Grass, Mountain, People, Tree
	Gate, Grass, Temple, Tree	People, Building, Flag, Parade, Sky	Water, Sky, People, Building, Boat	Building, Grass, Tree, Sky, People
	Jet, Mountain, Plane	Sky, Plane, Jet, Birds, Albatross	Sky, Plane, Jet, Birds, People	Sky, Plane, Jet, Tree, Clouds
	Athlete, Pool, Swimmer, Water	People, Pool, Swimmer, Water, Sky	Swimmer, Pool, Water, People, Sky	Swimmer, Pool, Water, People, Race
	Cat, Forest, Grass, Tiger	Water, Tree, Grass, Garden, Plants	Grass, Plants, Cat, Garden, Tree	Cat, Grass, Plants, Forest, Garden

Figure 8. Some sample images and the annotations by CMRM, ECMRM, and the proposed approach.

5. CONCLUSION

In this paper, we have proposed a method for AIA which is improved the result of ECMRM method. Instead of using the direct distribution of regions in ECMRM, we utilize the distribution of topics between regions. Moreover regional contexts are more accurate than textual contexts or the relationship between keywords in ECMRM. The proposed approach combines the regional contexts and visual topics for automatic image annotation, as for the dependence between these two kinds of information in the image, we use the MCDM approach based on TOPSIS method to integrate them. To obtain global distribution of topics, visual topics are learned from the training images by a PLSA approach. Furthermore, PLSA has used to model the regional contexts or the topics between regions. The proposed method is tested on a 5000 Corel data set and the results show that utilization of regional contexts or considering the relationship between regions improves the performance significantly and its incorporation with the visual topics leads to the best performance.

REFERENCES

[1] A Tousch, S Herbin, J Audibert. *Semantic hierarchies for image annotation: A survey*. Elsevier Ltd., Pattern Recognition. 2012; 40: 333-345.

[2] D Zhang, MdM Islam, G Lu. *A review on automatic image annotation techniques*. Elsevier Ltd., Pattern Recognition, Vol. 45, 2011.

[3] R Datta, D Joshi, J Li, JZ Wang. *Image Retrieval: Ideas, Influences, and Trends of the New Age*. ACM Comput. Surv. Vol. 40, No. 2, Article 5, 2008.

[4] AWM Smeulders, M Worring, S Santini, A Gupta, R Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Anal. Mach.Intell*. 2000; 22(12): 1349-1380.

[5] J Tang. Automatic Image Annotation and Object Detection. A thesis for the degree of Doctor of philosophy, University Of Southampton, 2008.

[6] Y Liu, D Zhang, G Lu, W Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*. 2007; 40: 262-282.

[7] JY Pan, HJ Yang, P Duygulu, Ch Faloutsos. *Automatic image captioning*. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). 2004; 1987-1990.

[8] JS Hare, PH Lewis. *Saliency-based models of image content and their application to auto-annotation by semantic propagation*. In Proceedings of Multimedia and the Semantic Web / European Semantic Web Conference. 2005.

[9] C Cusano, G Ciocca, R Schettini. *Image Annotation Using Svm*. In Proceedings of Internet Imaging IV, SPIE 5304. 2003; 5304: 330-338.

[10] E Chang, K Goh, G Sychay, G Wu. *CBSA: content based soft annotation for multimodal image retrieval using Bayes point machines*. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26-38, 2003.

[11] J Li, J Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*. 2003; 25(19): 1075-1088.

[12] J Tang, H Li, G Qi, T Chua. Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Trans. on Multimedia*. 2010; 12(2).

[13] J Liu, M Li, Q Liu, H Lu, S Ma. *Image annotation via graph learning*. Elsevier Ltd., Pattern Recognition. 2009; 42: 218-228.

[14] J Liu, B Wang, H Lu, S Ma. *A graph-based image annotation framework*. Pattern Recognition Letters. 2008; 29: 407-415.

[15] J Liu, M Li, W Ma, Q Liu, H Lu. *An Adaptive Graph Model For Automatic Image Annotation*. In Proceedings of the 8th ACM international workshop on Multimedia information retrieval. 2006.

[16] H Tong, J He, M Li, W Ma, HJ Zhang, C Zhang. Manifold-Ranking Based Keyword Propagation For Image Retrieval. *EURASIP J. Appl. Signal Process. Spec. Issue Inf. Min. Multimedia Database 21*. 2006: 1-10.

[17] D Zhou, O Bousquet, T Lal, J Weston, B Scholkopf. *Ranking On Data Manifolds*. In Proceedings of 18th Annual Conference on Neural Information Processing System. 2003: 169-176.

[18] S Abd manaf, MJ Nordin. *Review on Statistical Approaches for Automatic Image Annotation*. In International Conference on Electrical Engineering and Informatics, Selangor, Malaysia. August 2009.

[19] G Carneiro, AB Chan, PJ Moreno, N Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on PAMI*. 2007; 29(3).

[20] Y Mori, H Takahashi, R Oka. *Image-to-word transformation based on dividing and vector quantizing images with words*. In Proceedings of First International Workshop Multimedia Intelligent Storage and Retrieval Management. 1999.

[21] P Duygulu, K Barnard, J Freitas, D Forsyth. *Object recognition as machine translation: learning a lexicon for a fixed image vocabulary*. In Proceedings of the 7th European Conference on Computer Vision. 2002; 2353: 97-112.

[22] J Jeon, V Lavrenko, R Manmatha. *Automatic image annotation and retrieval using Cross-Media Relevance Model*. In Proceedings of the 26th annual international ACM SIGIR. 2003: 119-126.

[23] V Lavrenko, R Manmatha, J Jeon. *A model for learning the semantics of pictures*. In Proceedings of Advance in Neutral Information Processing. 2003.

[24] S Feng, R Manmatha, V Laverenko. Multiple Bernoulli Relevance Models for image and video annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2004: 1002-1009.
[25] Y Wang, T Mei, Sh Gong, XSh Hua. *Combining global, regional and contextual features for automatic image annotation*. Elsevier Ltd., Pattern Recognition. 2009; 42: 259-266.
[26] Zh Li, H Ma, Zh Shi, Zh Shi. A Probabilistic Model for Automatic Image Annotation and Retrieval. In *IEEE Ninth International Conference on Computer and Information Technology*. 2009.
[27] T Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*. 2001; 42: 177-196.
[28] ChL Hwang, K Paul Yoon. *Multiple Attribute Decision Making: Methods and Applications*. Springer-Verlag, New York. 1981.
[29] YJ Lai, TY Liu, ChL Hwang. TOPSIS for MODM. *European Journal of Operational Research*. 1994; 76(3): 486-500,.
[30] G Csurka, ChR Dance, L Fan, J Willamowski, C Bray. *Visual categorization with bags of keypoints*. In Proceedings of ECCV Workshop on Statistical Learning in Computer Vision. 2004: 1-16.
[31] DG Lowe. Distinctive image features from scale invariant keypoints. *Int. J. Comput. Vision*. 2004; 60(2): 91-110.
[32] R Xu, D Ii. Survey Of Clustering Algorithms. *IEEE Transactions On Neural Networks*. 2005; 16(3).
[33] L Fei Fei, P Perona. *A Bayesian Hierarchical Model for Learning Natural Scene Categories*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2005; 2: 524-531.
[34] Y Deng, BS Manjunath. Unsupervised Segmentation of Color-Texture Regions in Images and Video. *IEEE Trans. on PAMI*. 2001; 23(8): 800-810.
[35] S Kullback, RA Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951; 22(1): 79-86.
[36] AP Dempster, NM Laird, DB Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. 2007; 39(1): 1-38.
[37] K Hotta. *Scene Classification Based on Multi-resolution Orientation Histogram of Gabor Features*. In ICVS'08 Proceedings of the 6th international conference on Computer vision systems. 2008; 5008: 291-301.
[38] K Hotta. *Scene classification based on local autocorrelation of similarities with subspaces*. In 16th IEEE International Conference on Image Processing (ICIP). 2009: 2053-2056.
[39] K Hotta. *Object Categorization Based on Kernel Principal Component Analysis of Visual Words*. In Proceedings of IEEE Workshop on Application of Computer Vision. 2008: 1-8.
[40] K Grauman, T Darrell. *Discriminative Classification with Sets of Image Features*. In Proceeding of International Conference on Computer Vision. 2005: 1458-1465.
[41] S Lazebnik, C Schmid, J Ponce. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006: 2169-2178.
[42] D Comaniciu, V Ramesh, P Meer. Kernel-based object tracking. *IEEE Trans. on PAMI*. 2003; 25(5): 564-577.